Guest Lecture for 6.869 Advances in Computer Vision

Activity Recognition

Bolei Zhou
MIT CSAIL

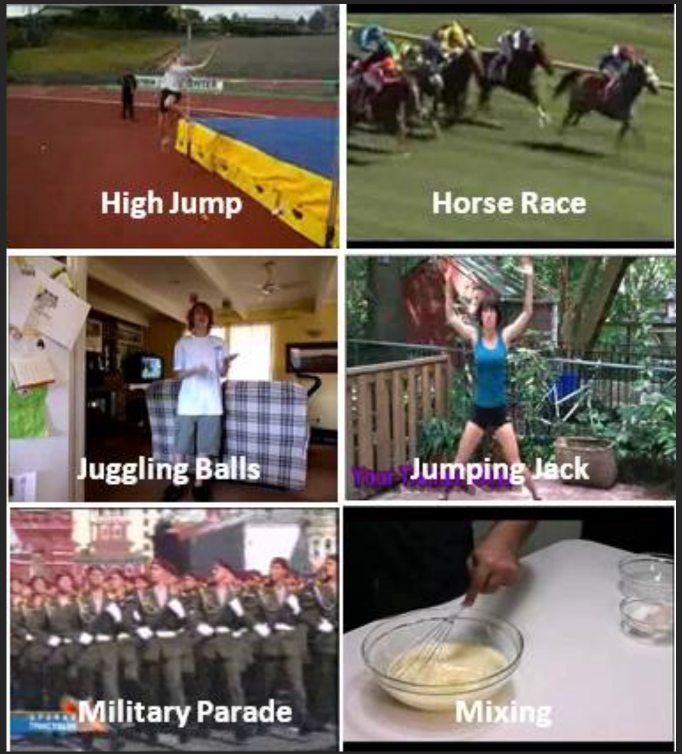
Challenge for Image Recognition

- Variation in appearance.

Challenge for Activity Recognition

- Describing activity at the proper level

Image recognition?
No motion needed?



Skeleton recognition?
Which activities?

Challenge for Activity Recognition

- Describing activity at the proper level

    A chain of events

Making chocolate cookies

Outline

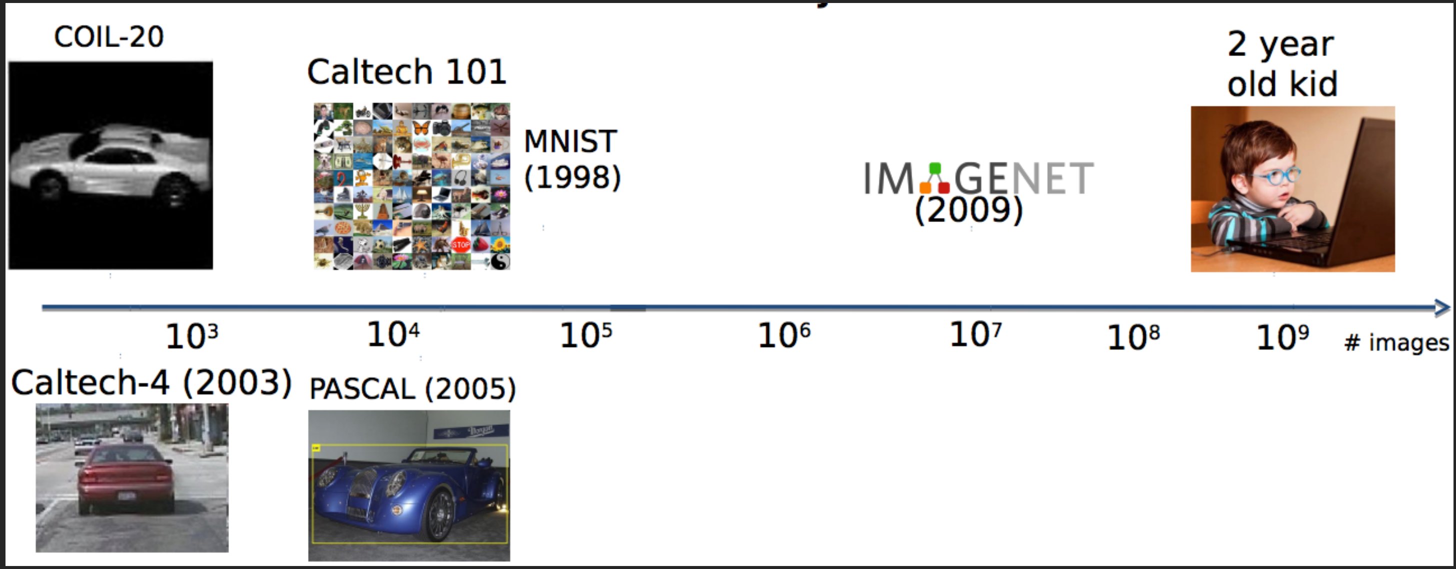
- Video Recognition Datasets
- Video Recognition Models

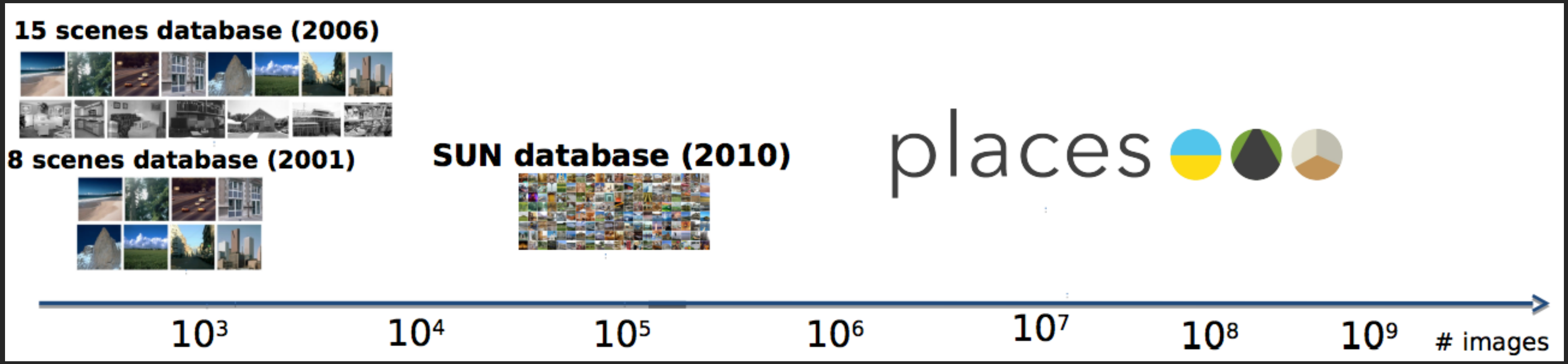
A little bit about my recent work:

- Temporal Relational Reasoning in Videos

Video Recognition Datasets

- Review on image datasets





Video Recognition Datasets

HMDB:51

Kinetics:400

KTH:6

UCF: 101

Moments: 339

ActivtyNet:200

Video Recognition Datasets

Two video collection methods:

- Collect videos from the web (Youtube, Flickr, etc)
- Crowd-sourcing video collection.

Video Recognition Datasets

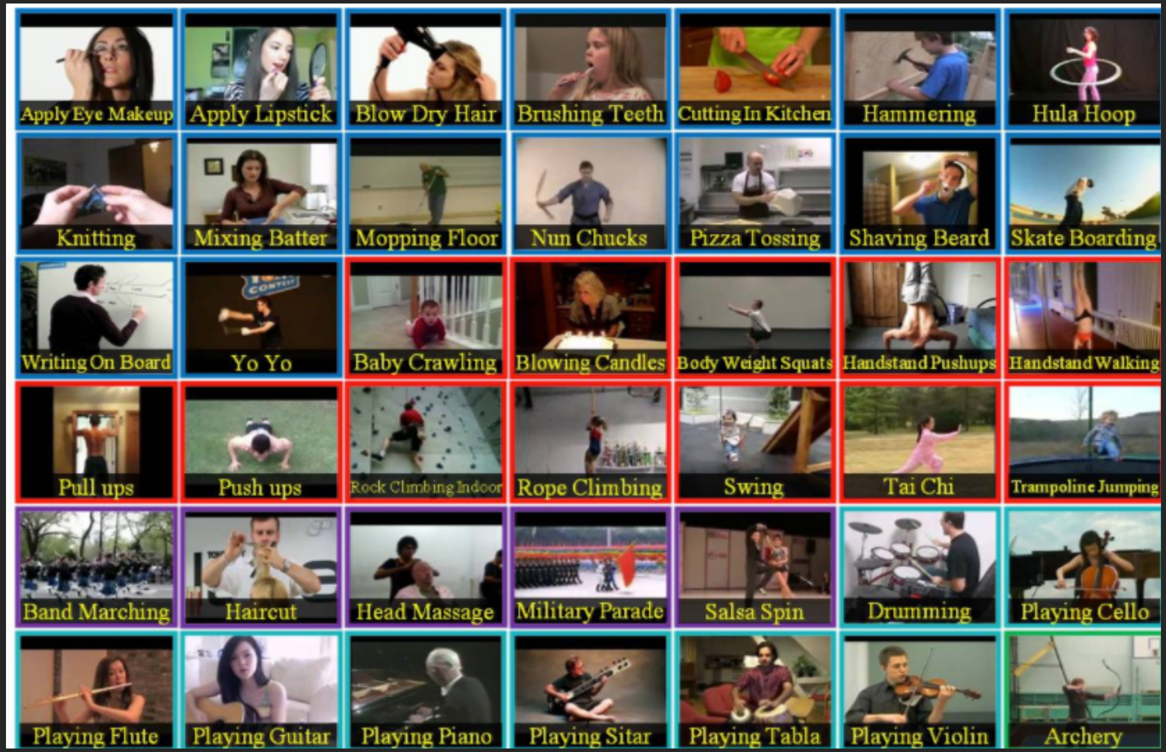
- KTH Dataset: recognition of human actions
- 6 classes, 2391 videos https://www.youtube.com/watch?v=Jm69kbCC17s

Recognizing Human Actions: A Local SVM Approach. ICPR 2004

Video Recognition Datasets

- UCF101 from University of Central Florida
- 101 classes, 9,511 videos in training

https://www.youtube.com/watch?v=hGhuUaxocIE

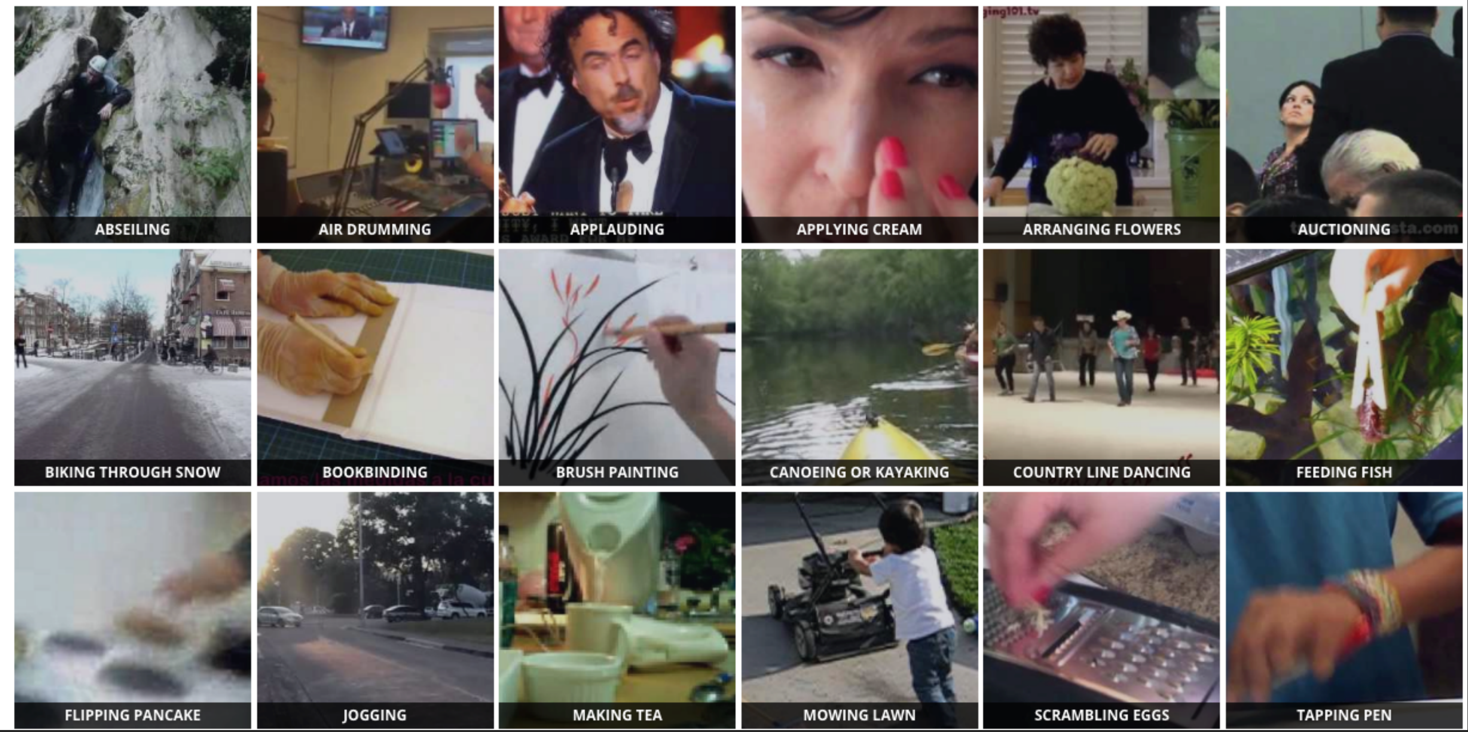


UCF101: A Dataset of 101 Human Action Classes From Videos in The Wild. 2012

Video Recognition Datasets

- Kinetics from Google DeepMind
- 400 classes, 239,956 videos in training

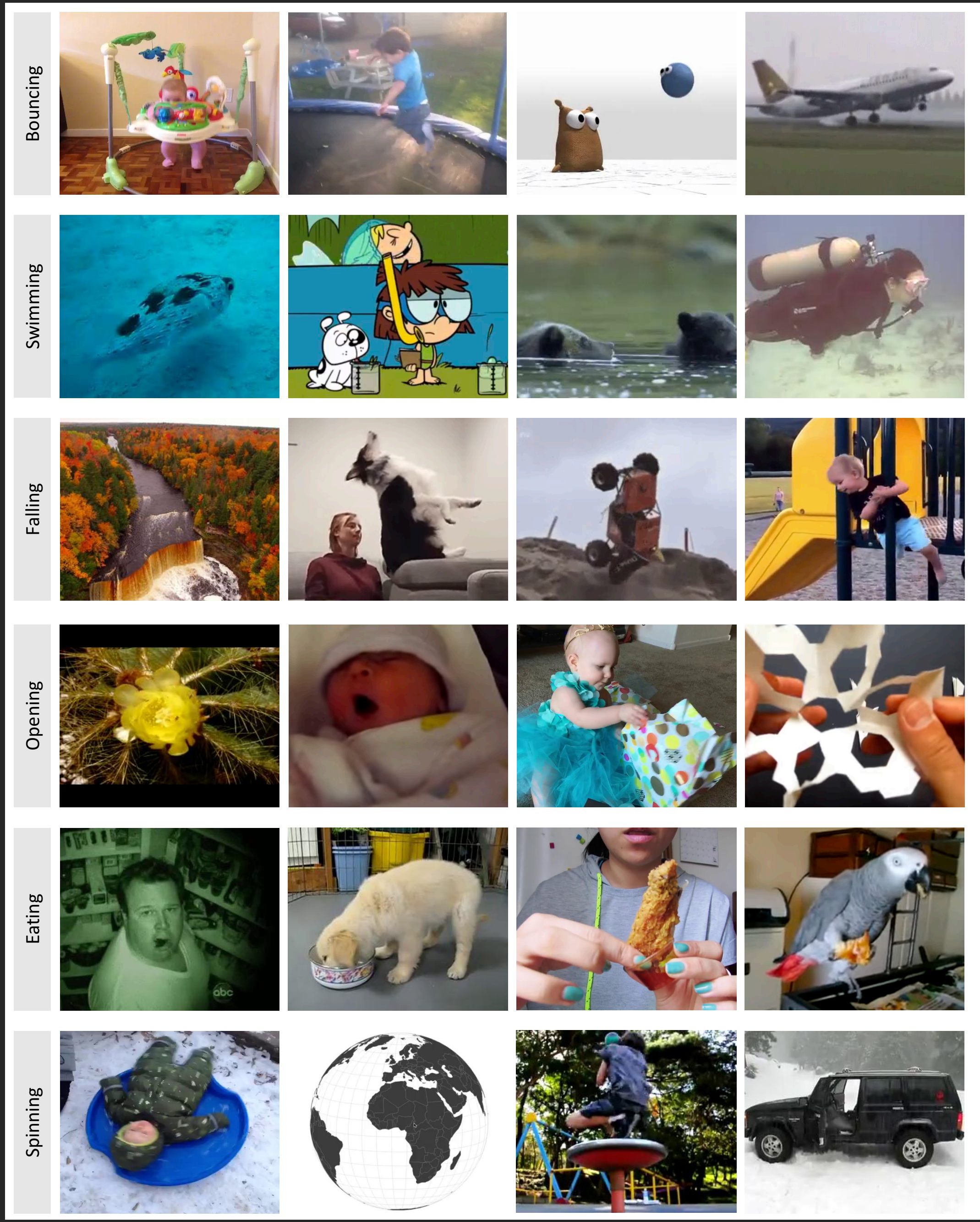
https://deepmind.com/research/open-source/open-source-datasets/kinetics/



Video Recognition Datasets

- Moments from MIT
- 1 million 3-second video from 339 generic actions

http://moments.csail.mit.edu/index_test.html



Video Recognition Datasets

- Charades dataset: Hollywood in Homes
- Crowdsourced video dataset

http://allenai.org/plato/charades/

Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. ECCV'16

Video Recognition Datasets

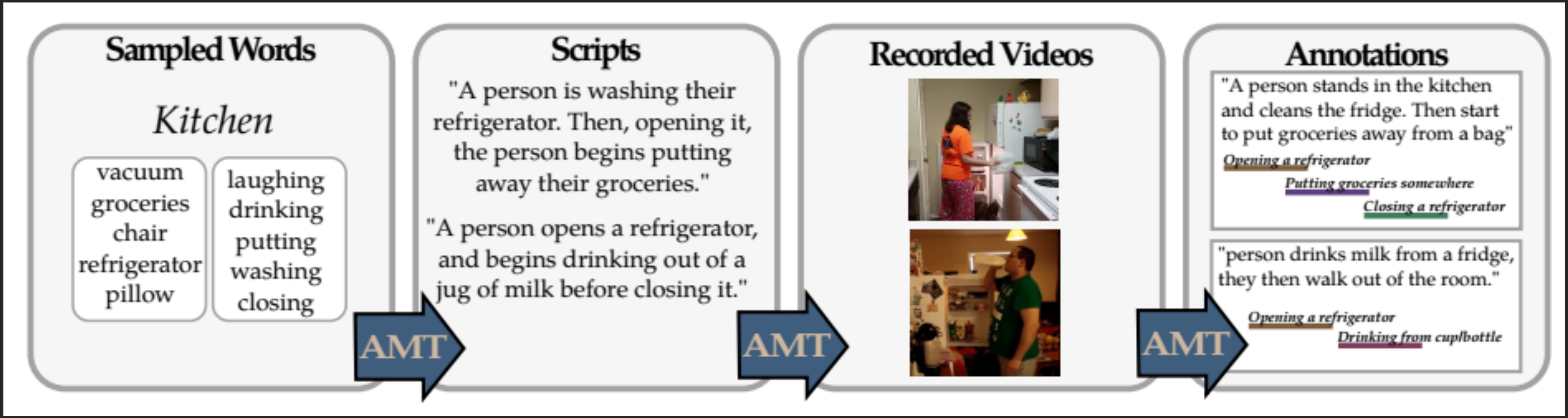
- Charades dataset: Hollywood in Homes
- Long chain of actions

https://www.youtube.com/watch?v=x9AhZLDkbyc

Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. ECCV'16

Video Recognition Datasets

- Charades dataset: Hollywood in Homes
- Crowd-sourced video dataset



Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. ECCV'16

Video Recognition Datasets

- Charades dataset: Hollywood in Homes
- Demo video

https://www.youtube.com/watch?v=x9AhZLDkbyc

Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. ECCV'16

Video Recognition Datasets

- Something-Something dataset: human object interaction
- 174 categories: 100,000 videos

- Holding something
- Turning something upside down
- Turning the camera left while filming something
- Opening something

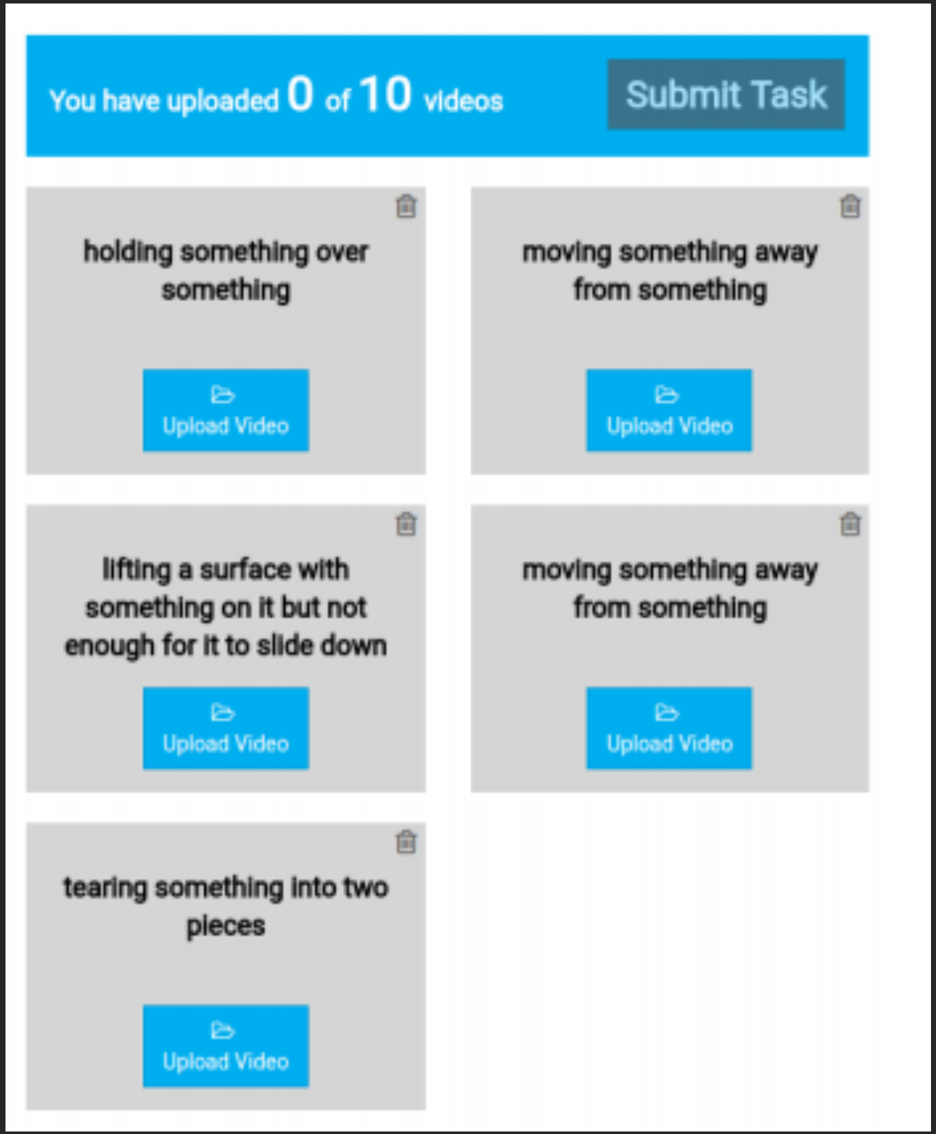
Poking a stack of something so the stack collapses

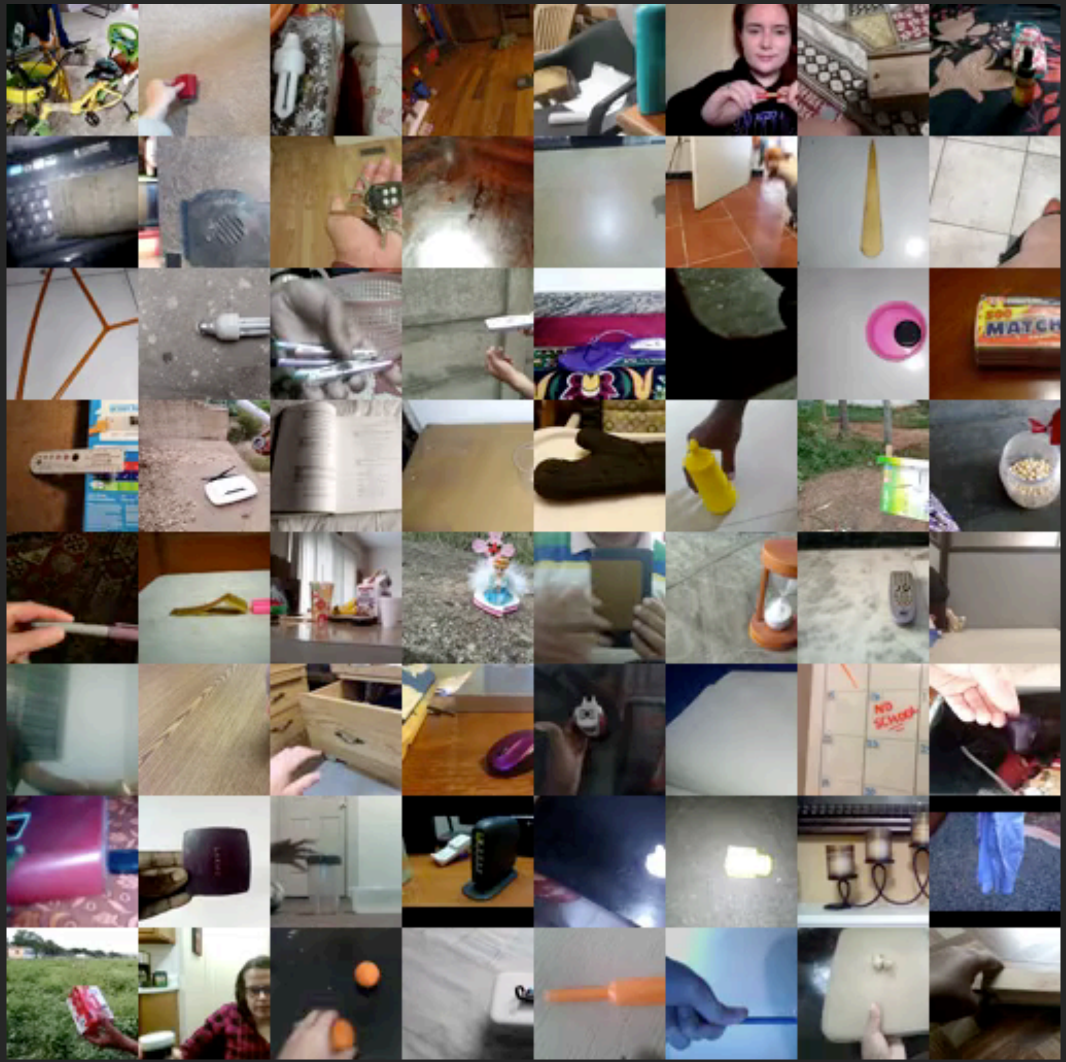
Plugging something into something

https://www.twentybn.com/datasets/something-something

Crowd-sourcing Video Collection

amazon mechanicalturk

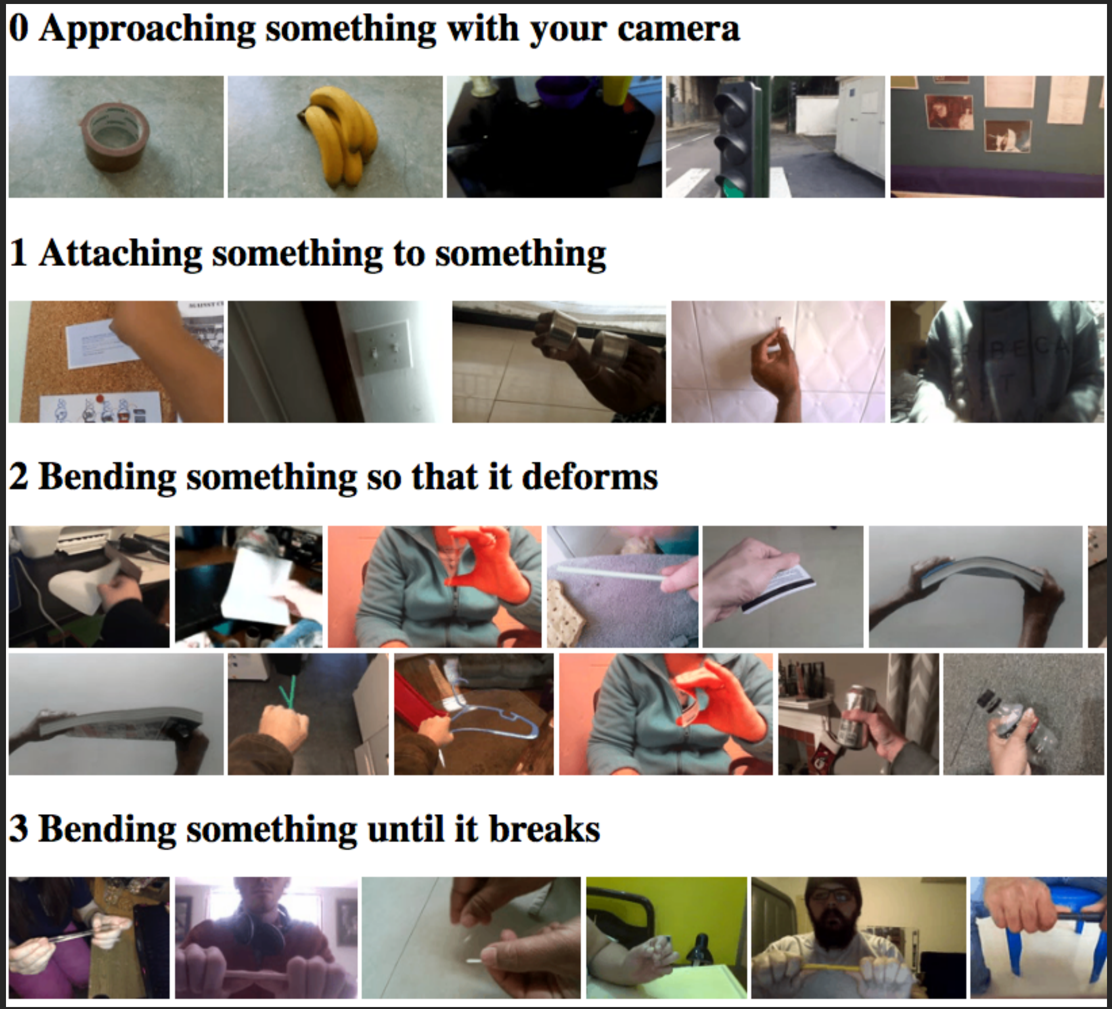




https://www.twentybn.com/datasets/something-something

Something-to-Something

http://visiongpu23.csail.mit.edu/deepscene/moments/models/datasets/something/plot_gif.html



Video = Sequence of RGB images

How to represent temporal information?

- Capture the temporal dependency
- Efficiency: 1min 25fps video = 1500 images

Video Recognition Models

- Pre-Deep learning era

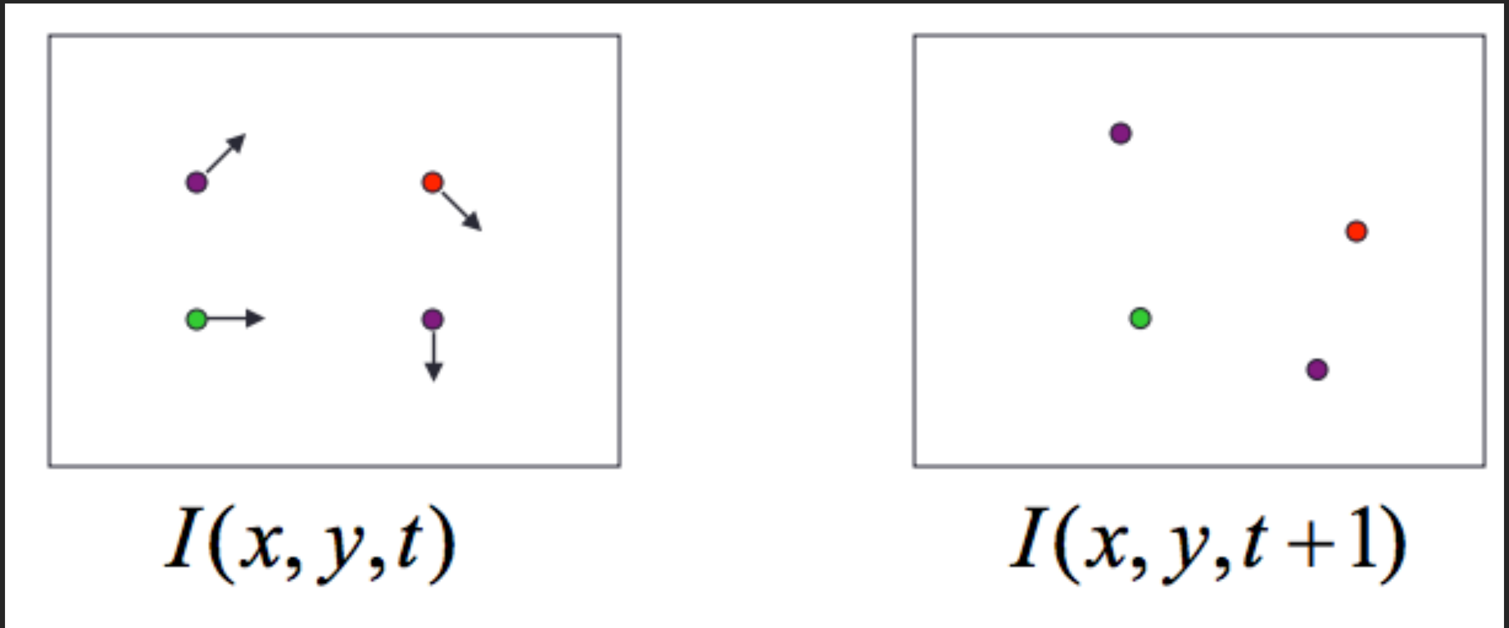
Optic flow, trajectories, bag of words.

- Deep learning era

Neural Networks

Pre-deep learning Activity Recognition

- Optic Flow: the displacement of pixels
- Gesture lecture by Ce Liu next week on motion estimation



Motion Representations in Activity Recognition

- Optic Flow

https://www.youtube.com/watch?v=JSzUdVBmQP4

Motion Representations in Activity Recognition

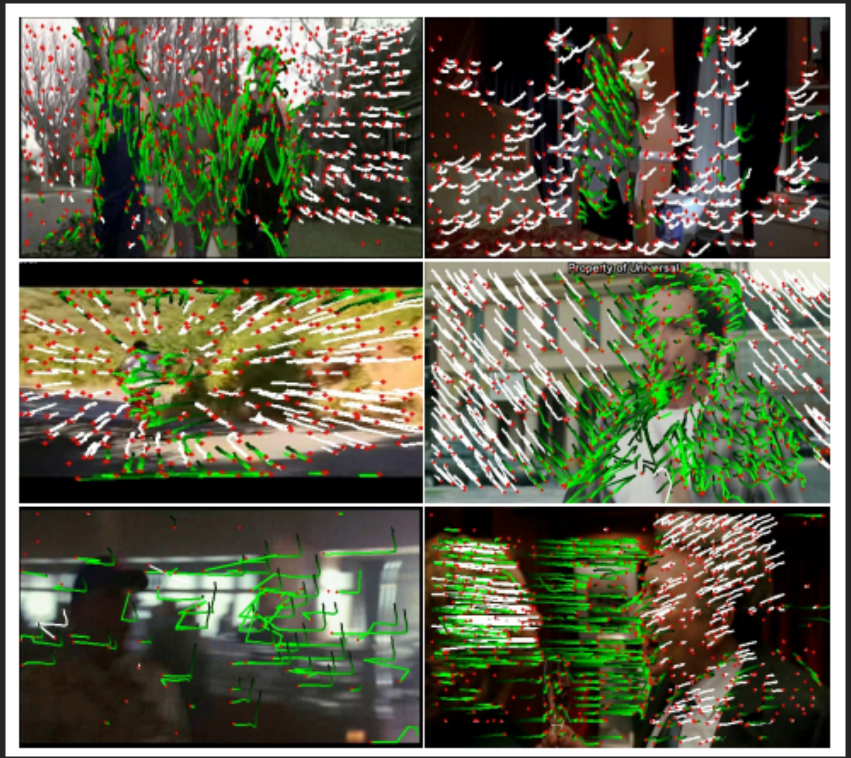
- Trajectories: key-point tracking over frames

https://www.youtube.com/watch?v=YN2lDqXz-uc

Motion Representations in Activity Recognition

Improved Dense Trajectory (iDT)

- Global motion compensation (camera motion removal)
- Features from trajectories and HoG
- Bag of trajectories + Fisher Vector + PCA

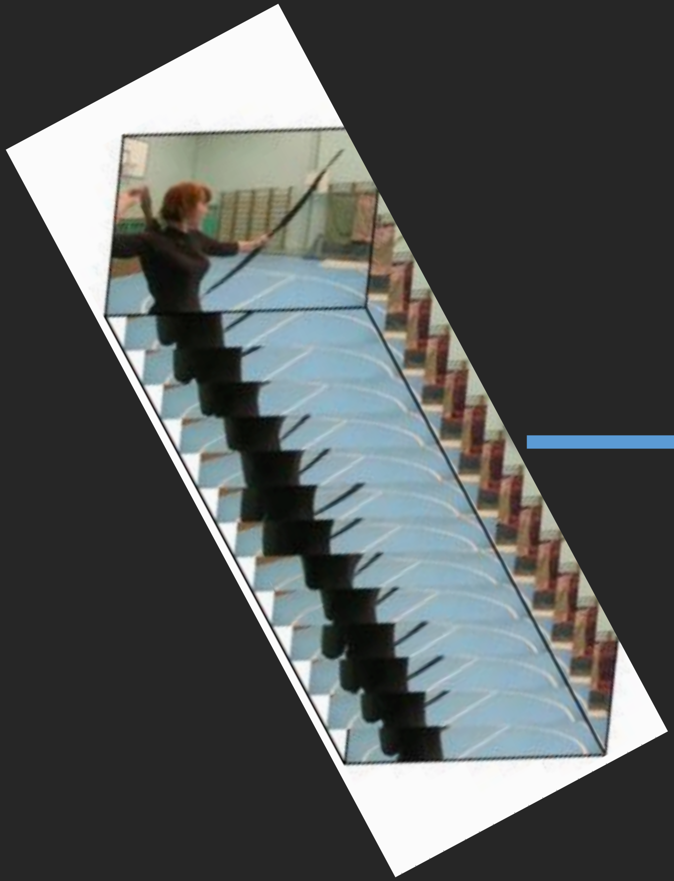


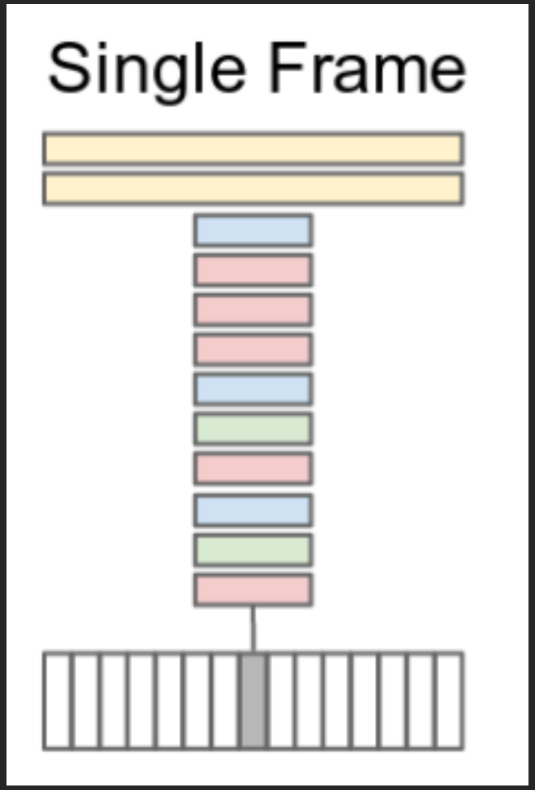
Action Recognition with Improved Trajectories. ICCV 2013

Deep Learning Models for Activity Recognition

- RGB frame fusion network
- 2-stream network
- 3D convolution network
- Temporal segment network

Single-frame image model

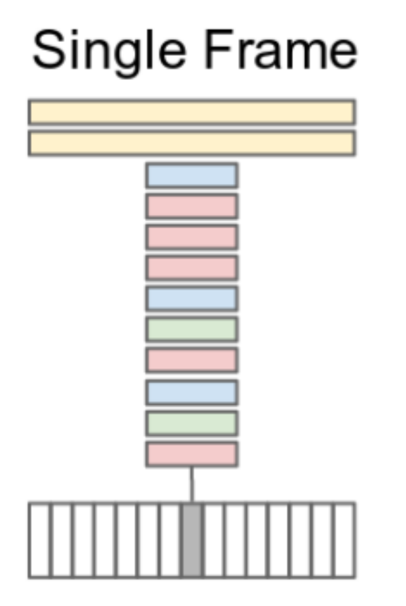




Large-scale Video Classification with Convolutional Neural Networks, CVPR 2014

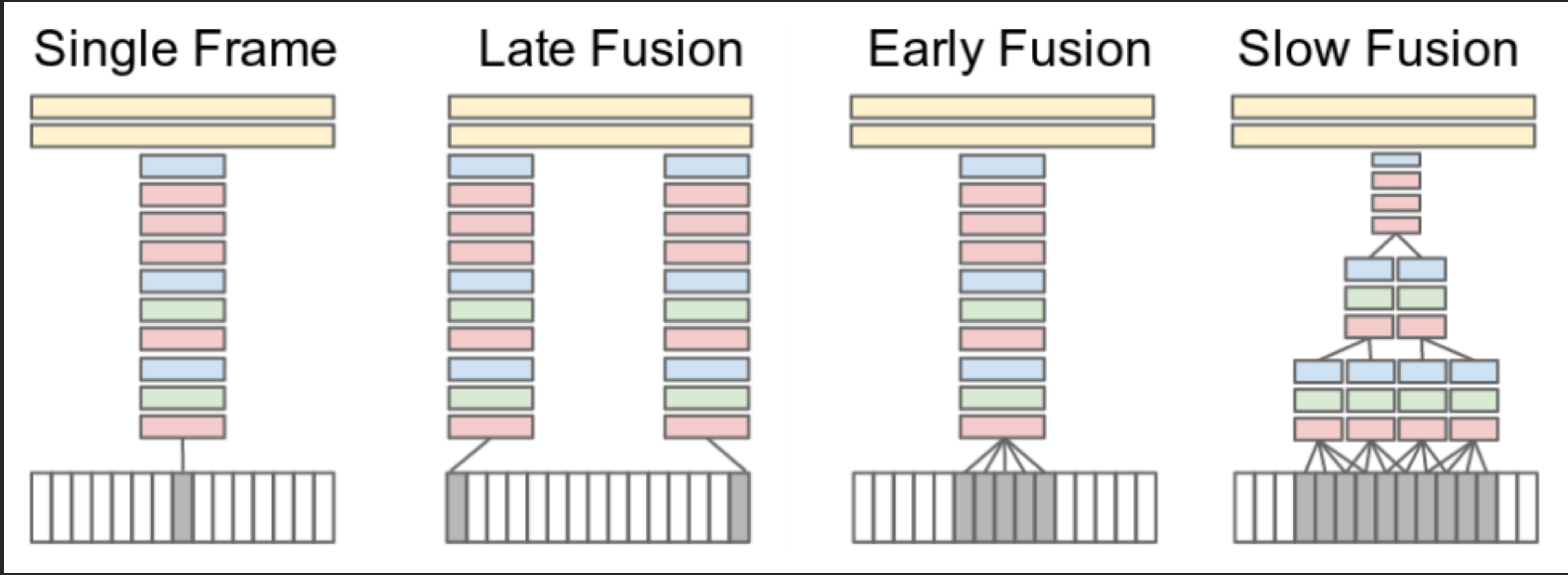
Performance on the UCF101

Spatial ConvNets	Temporal ConvNets	Two-Stream
72.7%	81.0%	87.0%



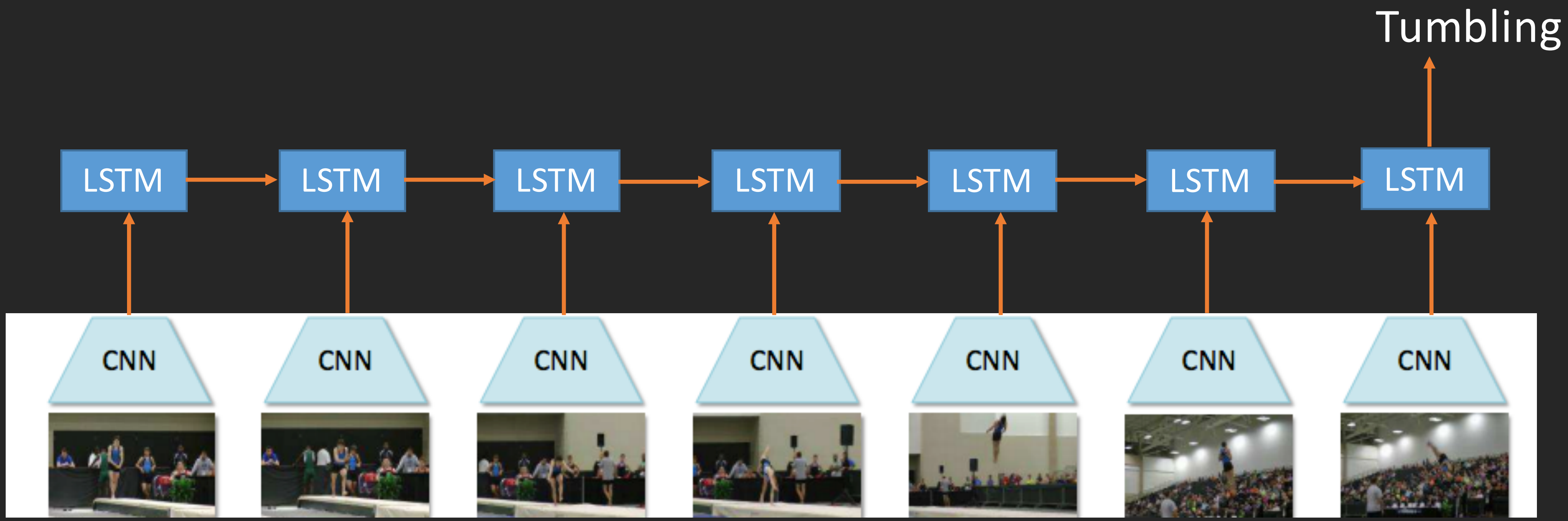
Large-scale Video Classification with Convolutional Neural Networks, CVPR 2014

Multi-frame fusion model



Large-scale Video Classification with Convolutional Neural Networks, CVPR 2014

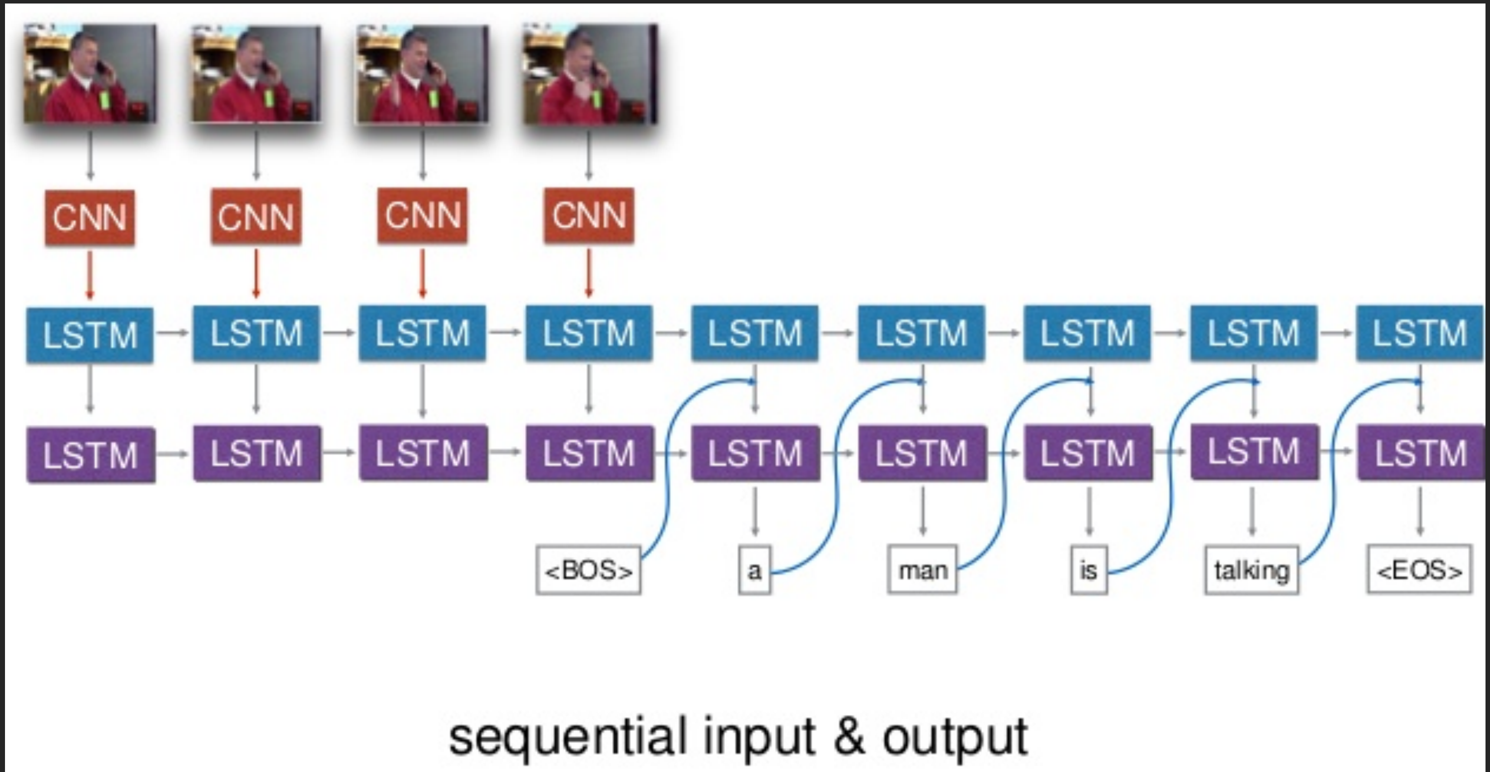
Multi-frame LSTM fusion model



Long-term Recurrent Convolutional Networks for Visual Recognition and Description. CVPR 2015

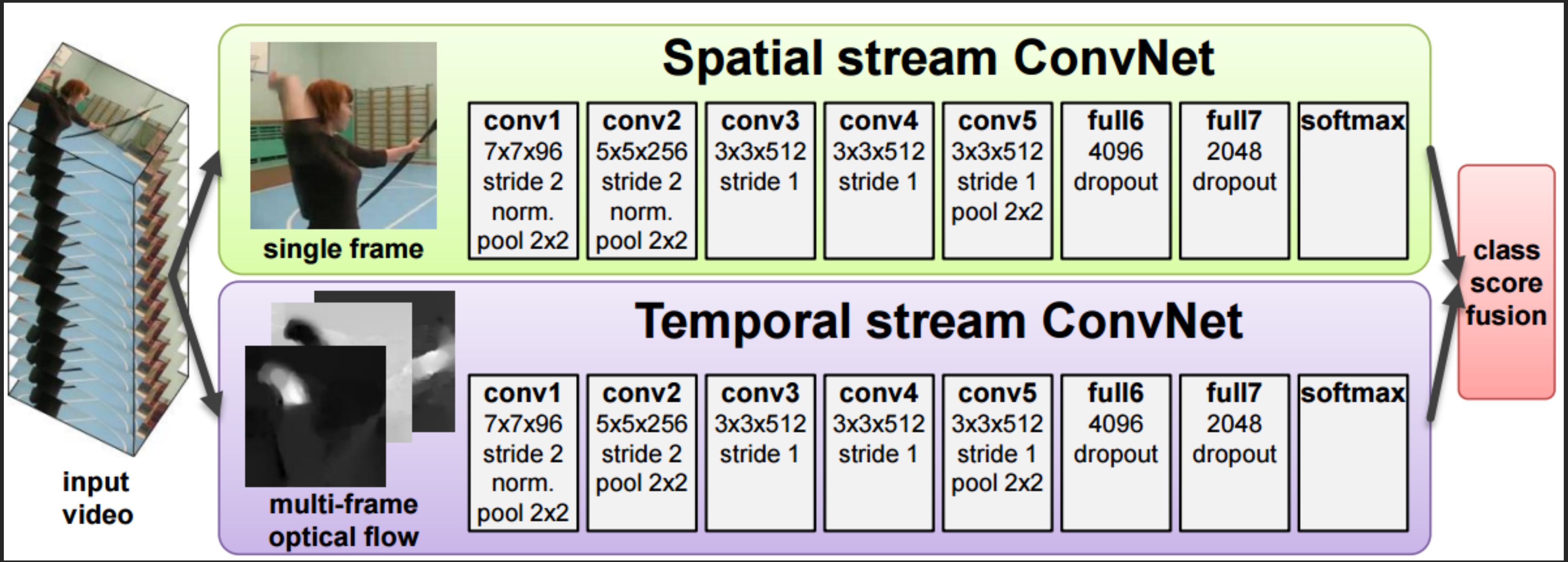
LSTM: recursive neural networks

- Video Captioning



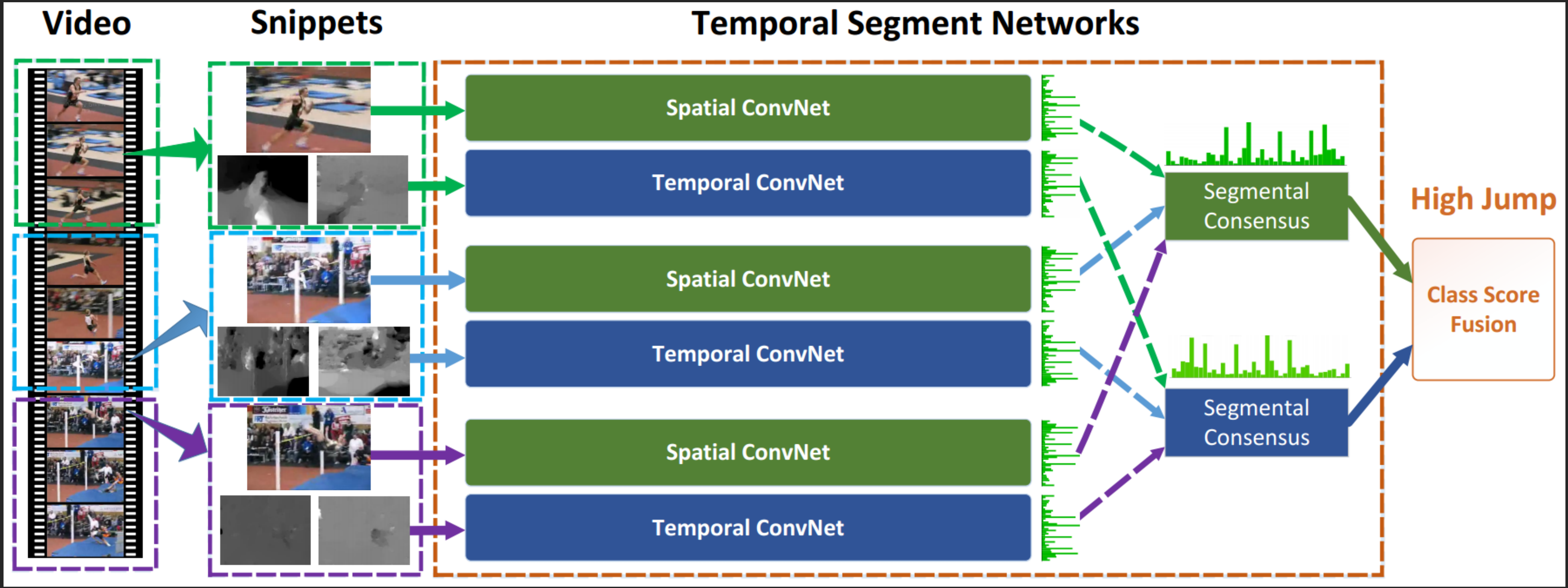
Sequence to Sequence – Video to Text https://arxiv.org/pdf/1505.00487.pdf 2015

2-Stream Network



Two-Stream Convolutional Networks for Action Recognition in Videos, NIPS 2014

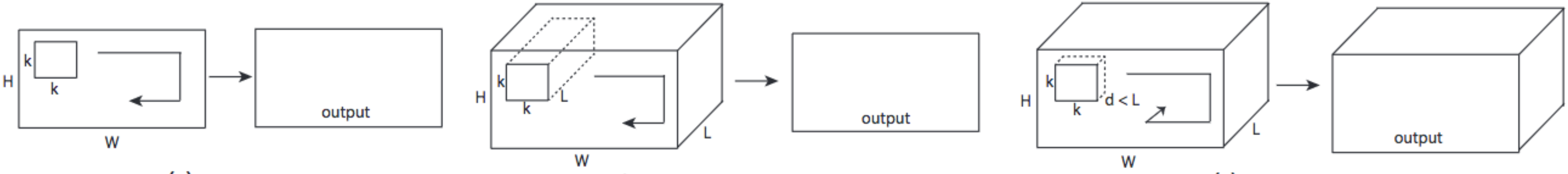
Temporal segment network



Temporal Segment Networks: Towards Good Practices for Deep Action Recognition, ECCV 2016

3D convolutional Networks

Computationally expensive, and a lot of model parameters



(a) 2D convolution (b) 2D convolution on multiple frames (c) 3D convolution

If it is RGB frame rather than grey frame, it is actually 4D convolution.

H x W x C x T

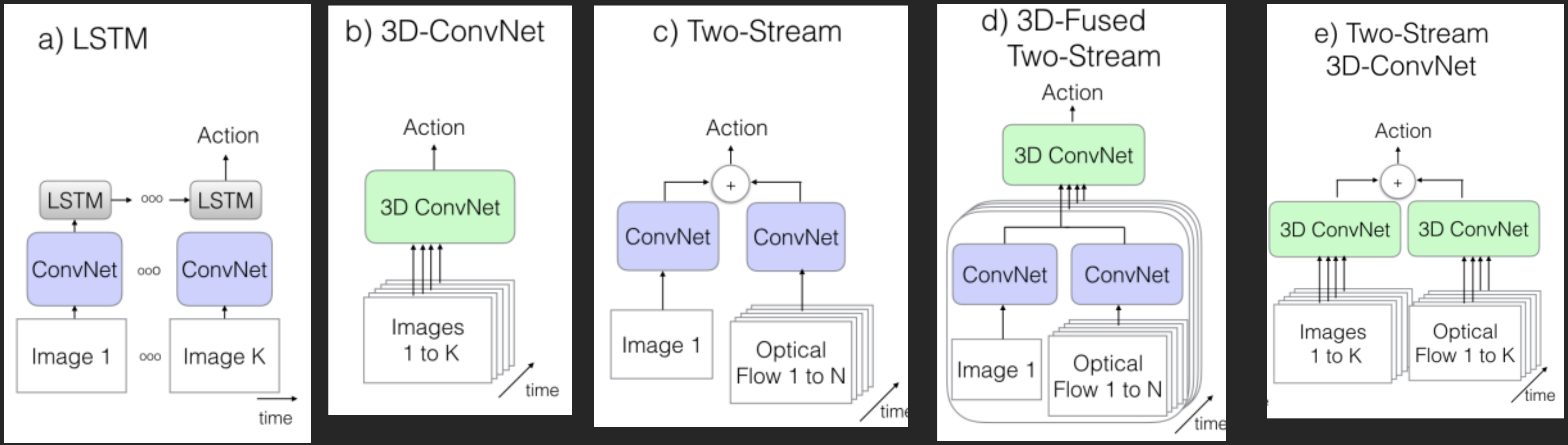
Learning Spatiotemporal Features with 3D Convolutional Networks, ICCV 2015

3D convolutional Networks

- 3D filters at the first layer.

Learning Spatiotemporal Features with 3D Convolutional Networks, ICCV 2015

Summary of Video Recognition Networks



Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. ICCV 2017

Pose Estimation in Videos

https://github.com/ZheC/Realtime_Multi-Person_Pose_Estimation

Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. CVPR'17

Pose Estimation in Videos

Demo Video:
https://www.youtube.com/watch?v=pW6nZXeWlGM&t=77s

https://github.com/ZheC/Realtime_Multi-Person_Pose_Estimation

Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. CVPR'17

Some of my latest work:

Temporal Relational Reasoning in Videos

Bolei Zhou, Alex Andonian, Antonio Torralba

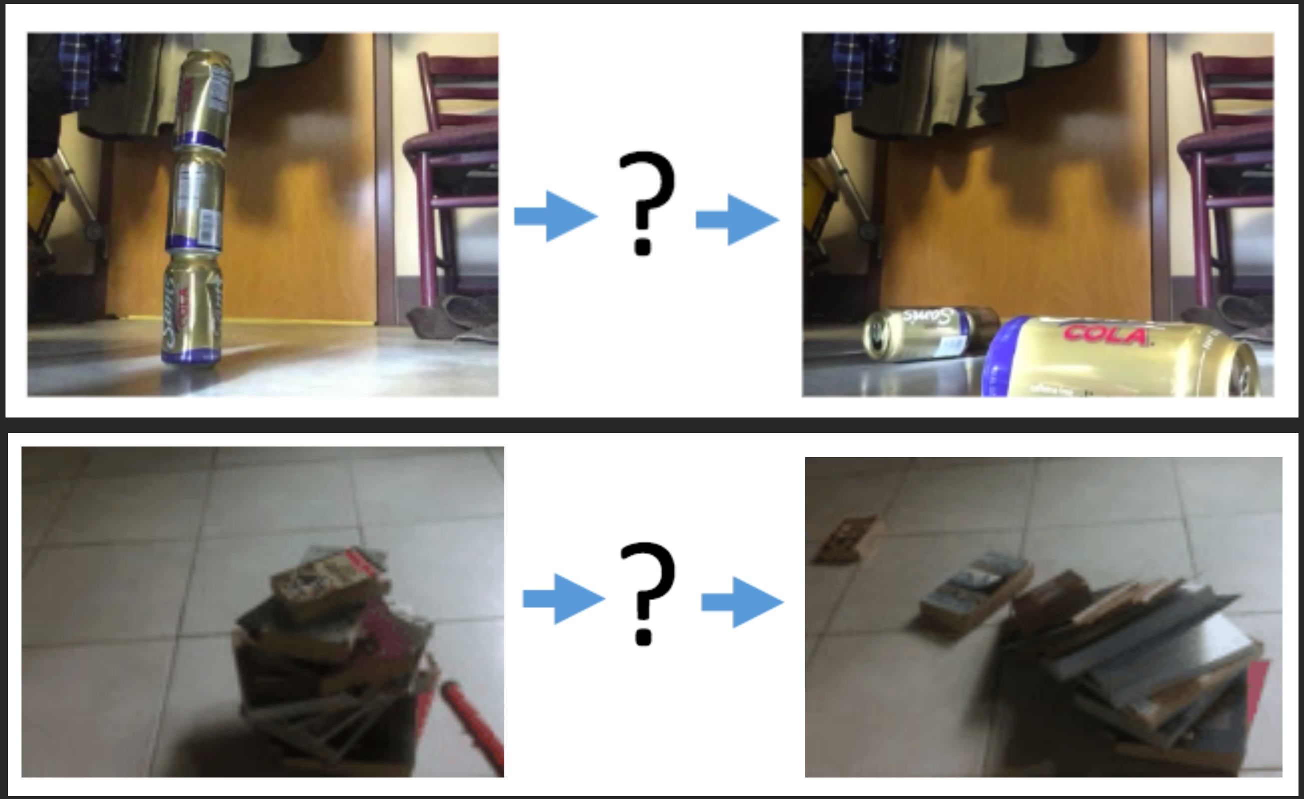
CVPR'18 submission

https://arxiv.org/pdf/1711.08496.pdf

Temporal Relational Reasoning

- Infer the temporal relation between frames.

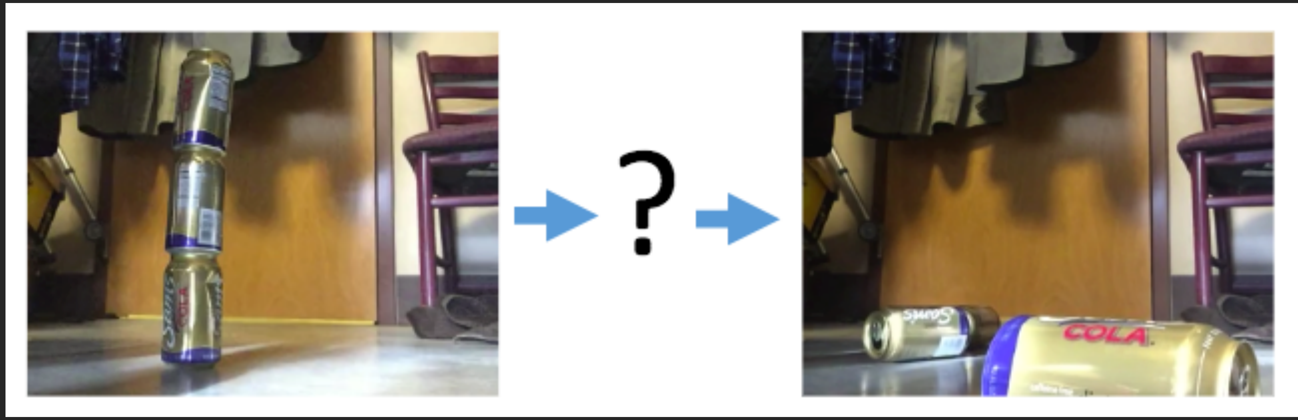
Poking a stack of something so it collapses

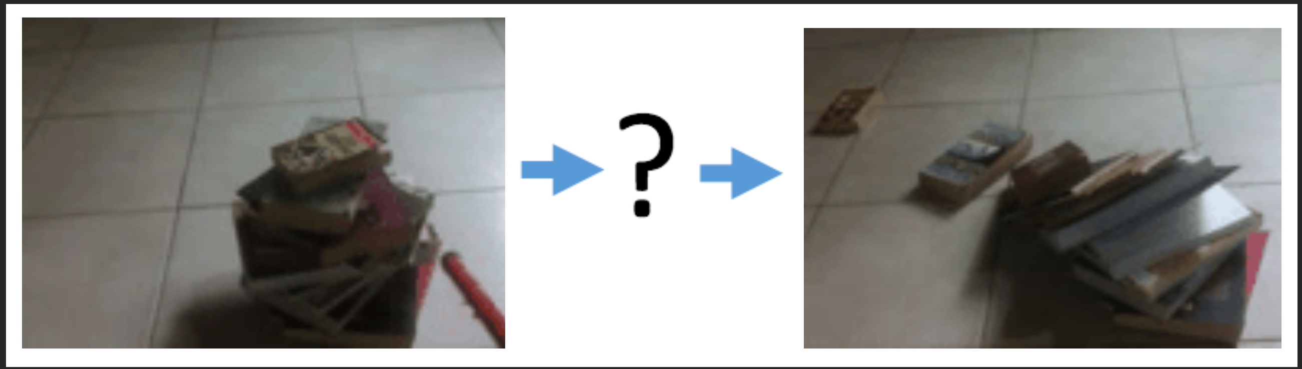


Temporal Relational Reasoning

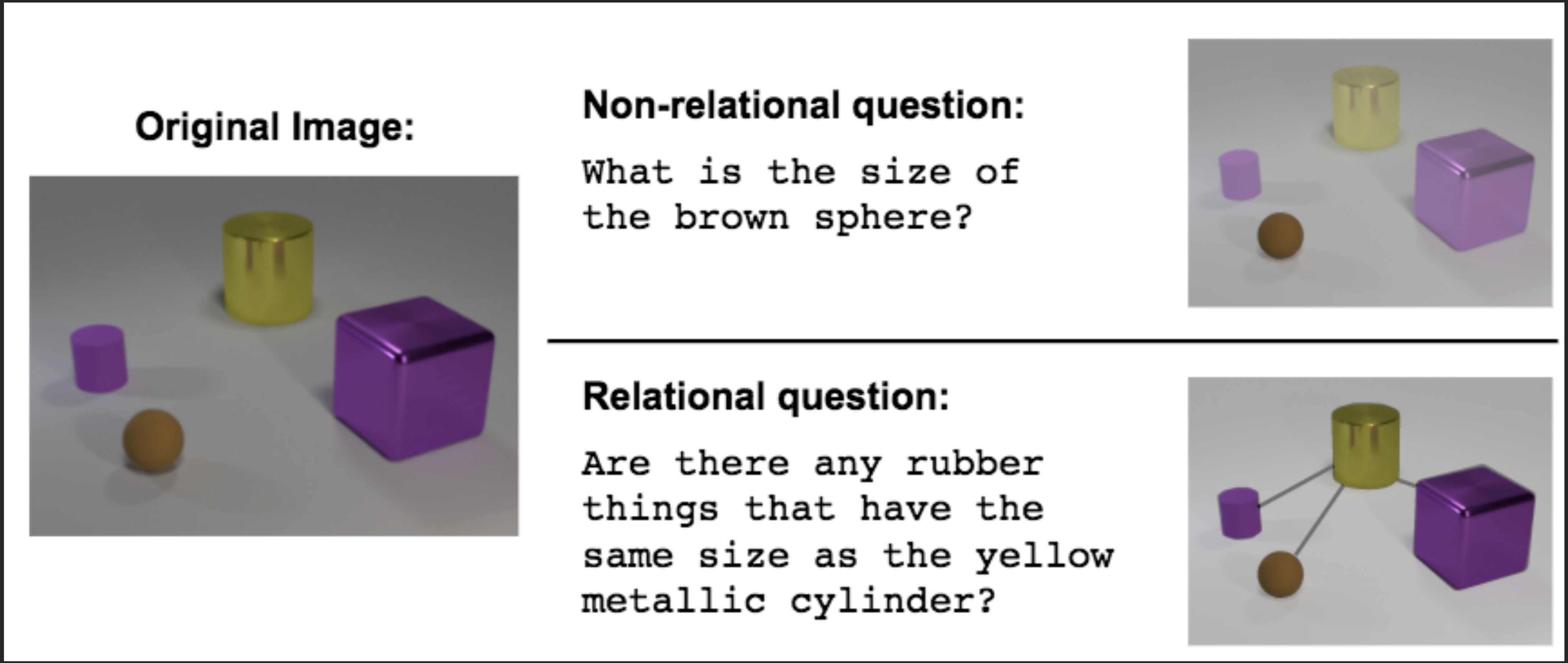
- It is the **temporal transformation/relation** that defines the activity, rather than the **appearance of objects**.

Poking a stack of something so it collapses

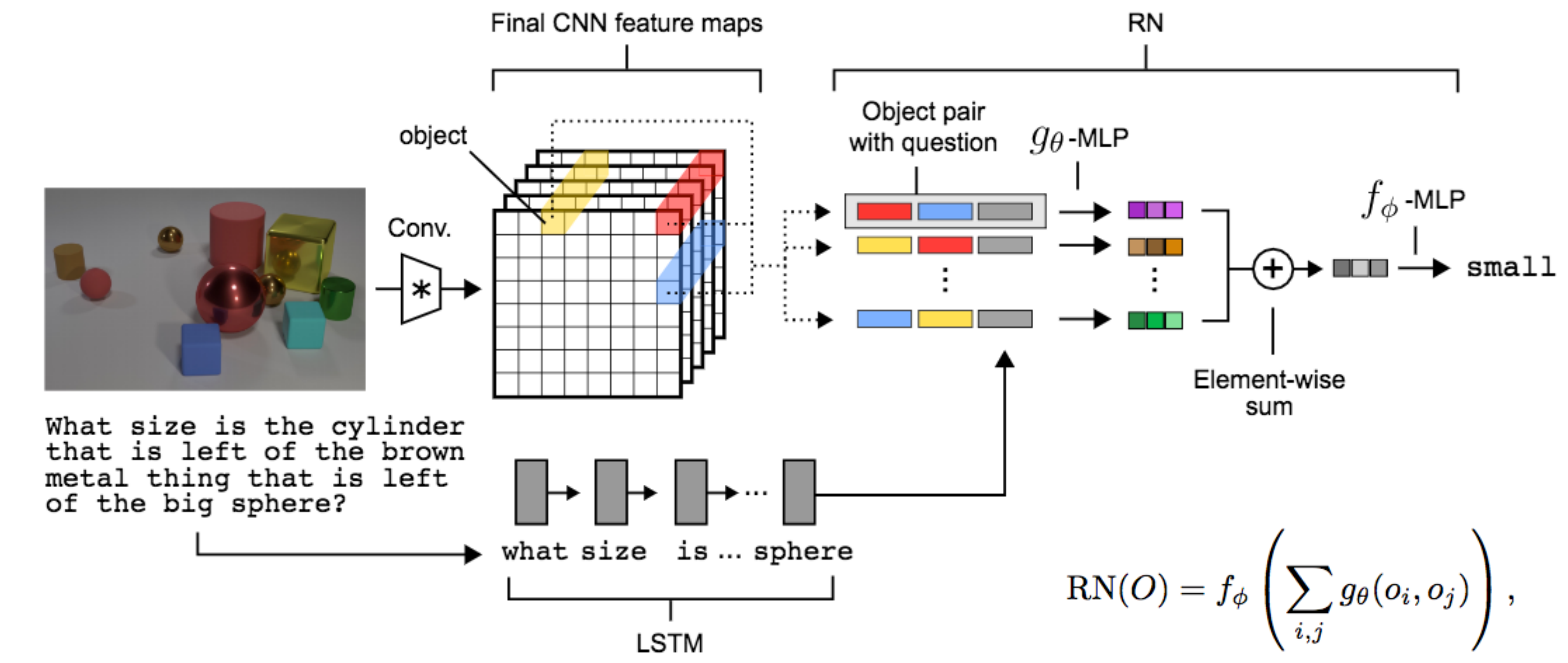




Relational Reasoning for Visual Question Answering



Relational Reasoning for Visual Question Answering

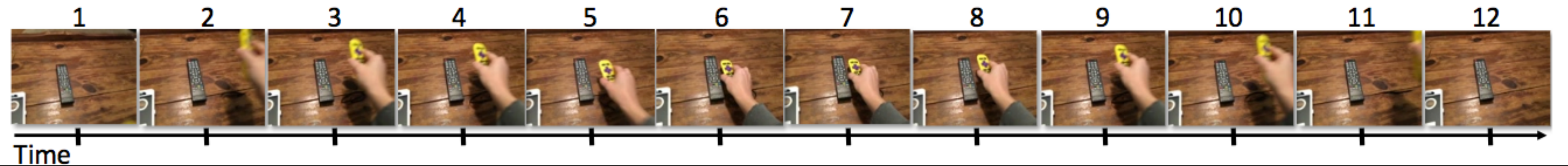


$$\mathrm{RN}(O) = f_\phi\left(\sum_{i,j} g_\theta(o_i, o_j)\right),$$

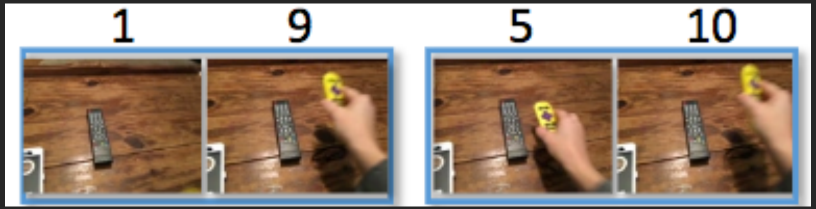
Google DeepMind: A simple neural network modulefor relational reasoning. https://arxiv.org/pdf/1706.01427.pdf

Temporal Relations in Videos

Pretending to put something next to something



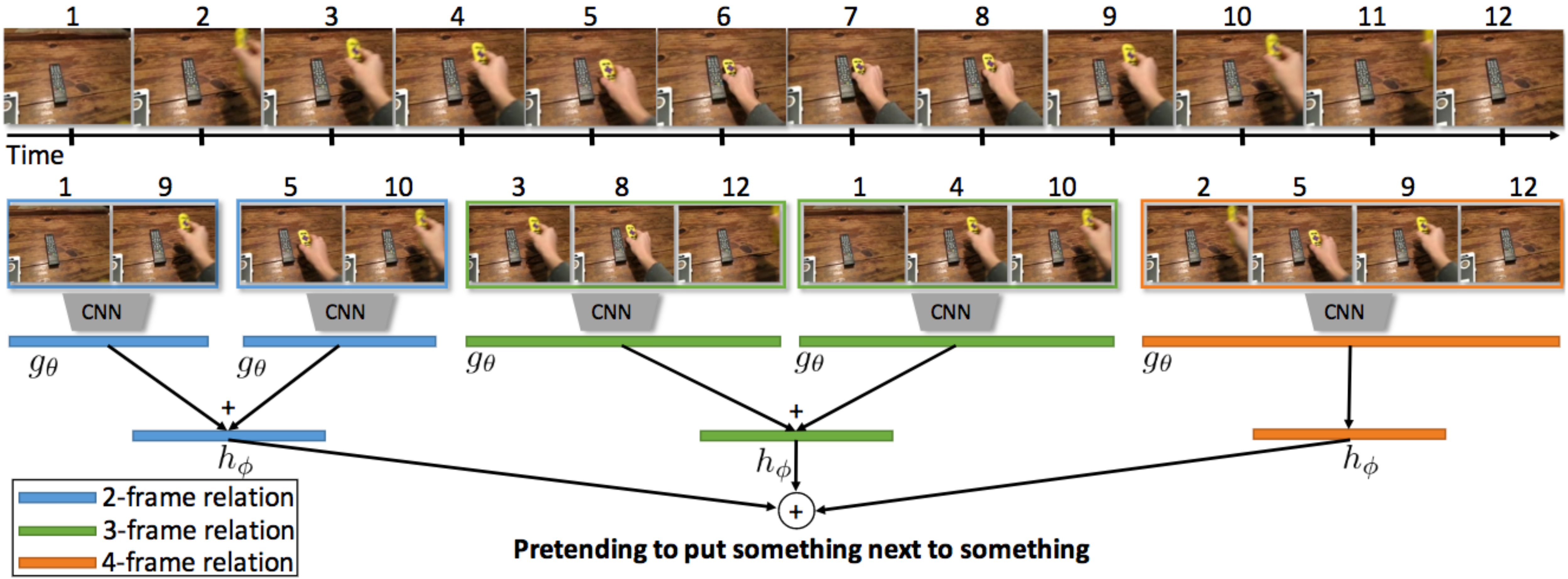
2-frame relations



3-frame relations

4-frame relations

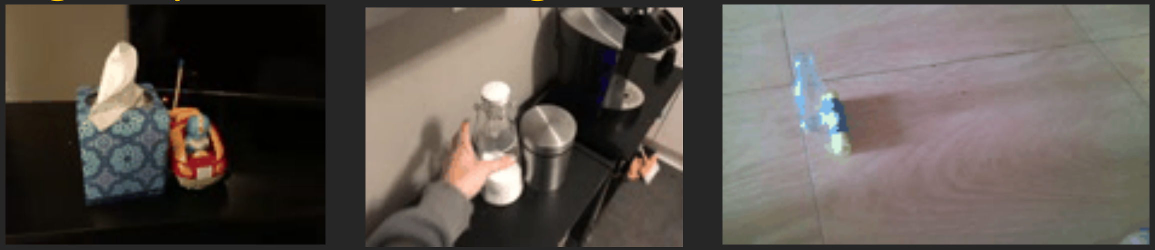
Framework of Temporal Relation Networks



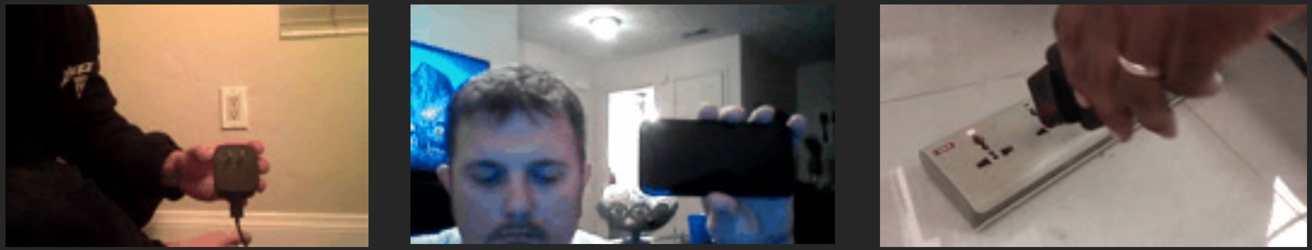
Something-Something Dataset

- 100 K videos from 174 human-object interaction classes.

Moving something away from something



Plugging something into something

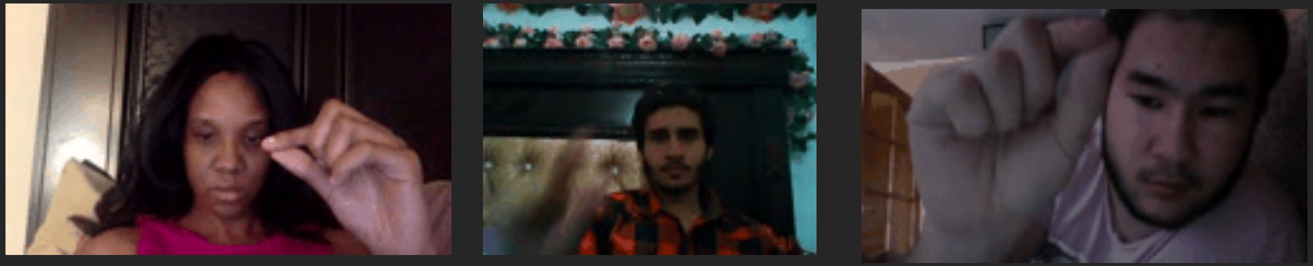


Pulling two ends of something so that it gets stretched

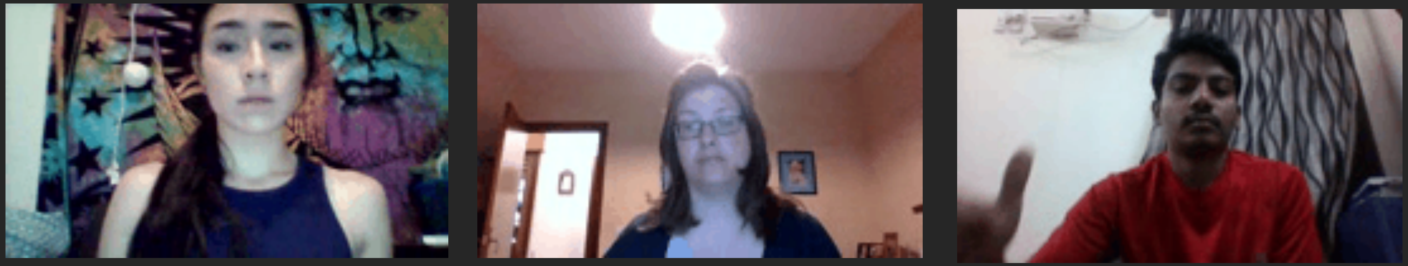
Jester Dataset

- 140 K videos from 27 gesture classes.

Zooming in with two fingers



Thumb down



Drumming fingers

Experimental Results

- On Something-Something dataset

model	Top1 acc.(%)	Top5 acc.
single frame	11.41	33.39
2-frame TRN	22.23	48.80
3-frame TRN	26.22	54.15
4-frame TRN	29.83	58.21
5-frame TRN	30.39	58.29
7-frame TRN	31.01	59.24
MultiScale TRN	33.01	61.27
MultiScale TRN (10-crop)	**34.44**	**63.20**

model	Top1 acc.(%)
Yana Hasson	25.55
Harrison.AI	26.38
I3D by [8]	27.23
Guillaume Berger	30.48
Besnet (Top1 on leaderboard)	31.66
MultiScale TRN	**33.60**

Experimental Results

- On Jester dataset

model	Top1 acc.(%)	Top5 acc.
single frame	63.60	92.44
2-frame TRN	75.65	94.40
MultiScale TRN	93.70	99.59
MultiScale TRN (10-crop)	**95.31**	**99.86**

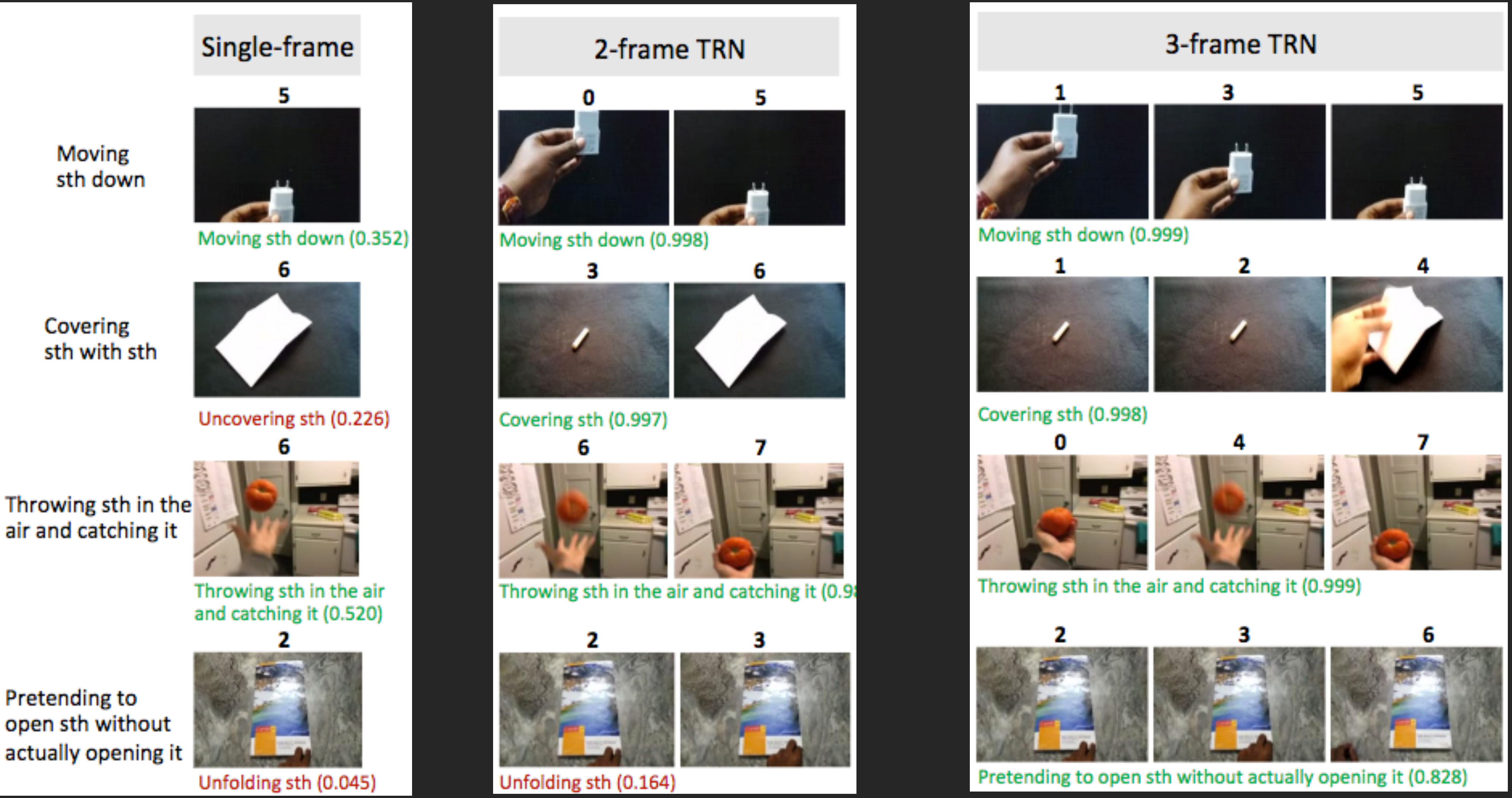
model	Top1 acc.(%)
20BN's Jester System	82.34
VideoLSTM	85.86
Guillaume Berger	93.87
Ford's Gesture Recognition System	94.11
Besnet (Top1 on leaderboard)	94.23
MultiScale TRN	**94.78**

Experimental Results

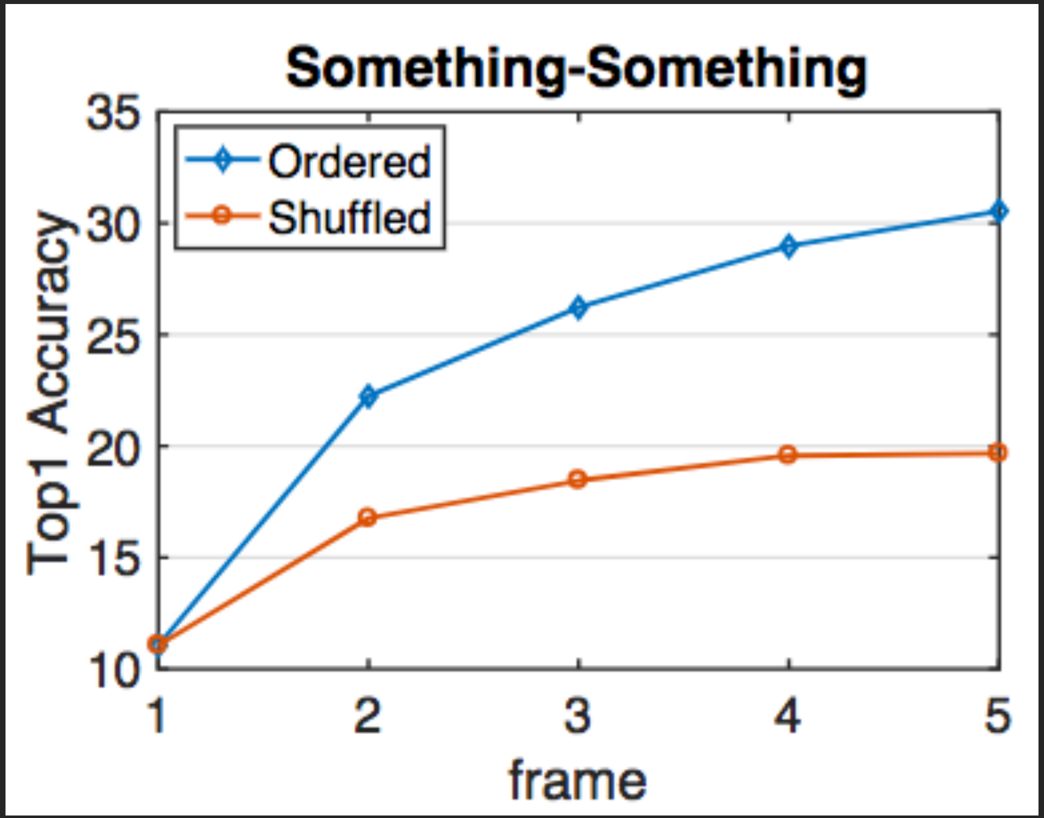
- Demo Video:

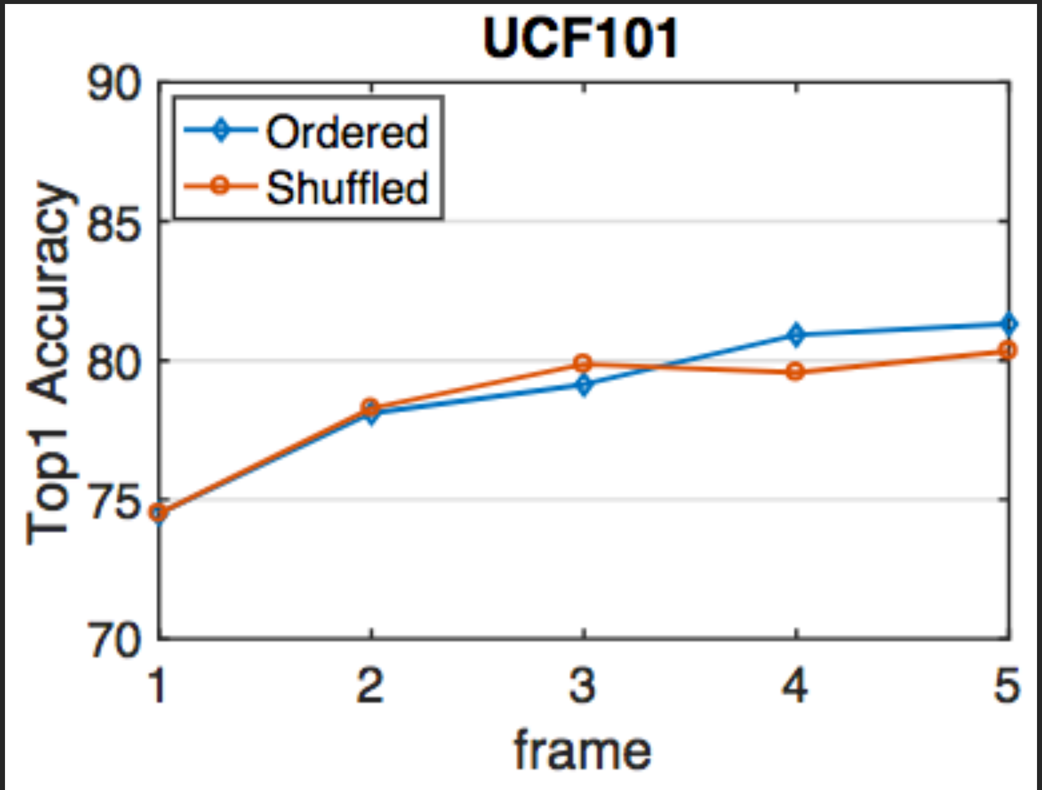
http://relation.csail.mit.edu/

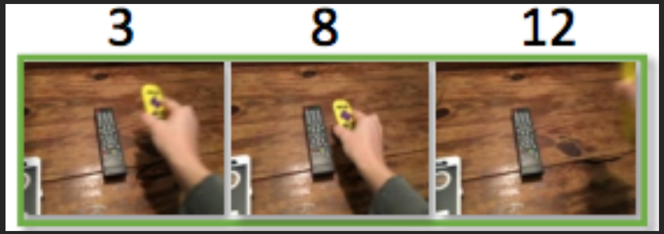
Common sense knowledge learned by models



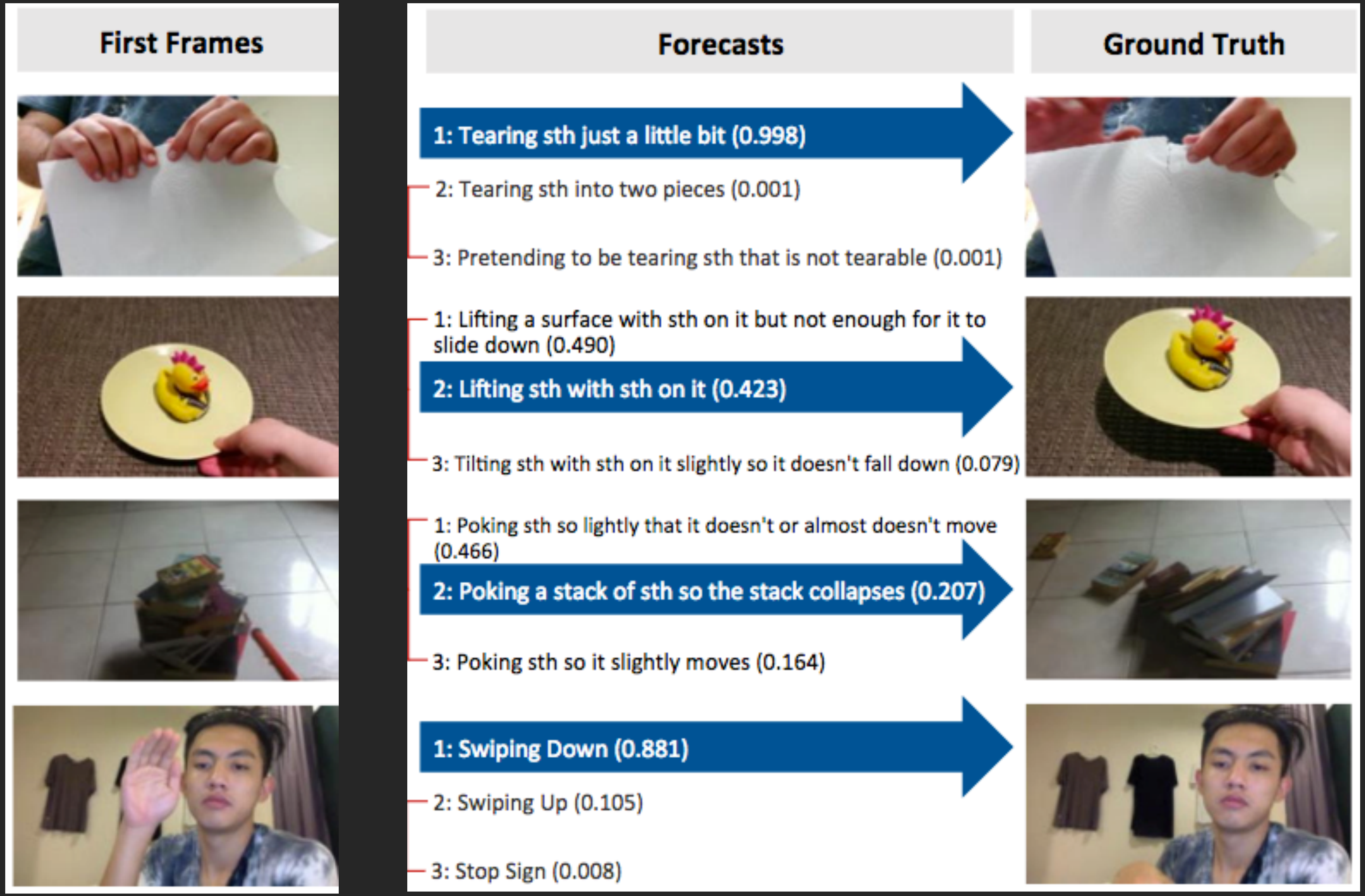
Importance of temporal orders







Activity Forecasting

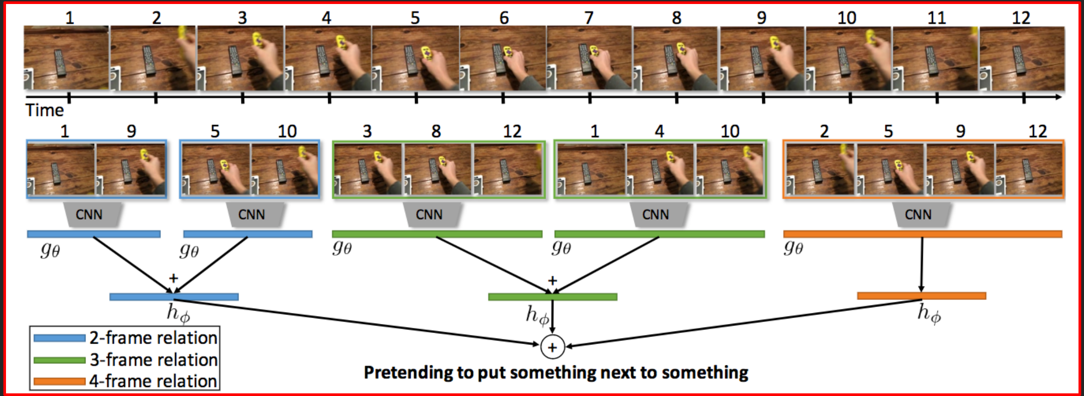


Activity Forecasting

	Something		Jester	
Data	baseline	TRN	baseline	TRN
first 25%	9.08	11.14	27.25	34.23
first 50%	10.10	19.10	41.43	78.42
full	11.41	33.01	63.60	93.70

Future Directions in Activity Recognition

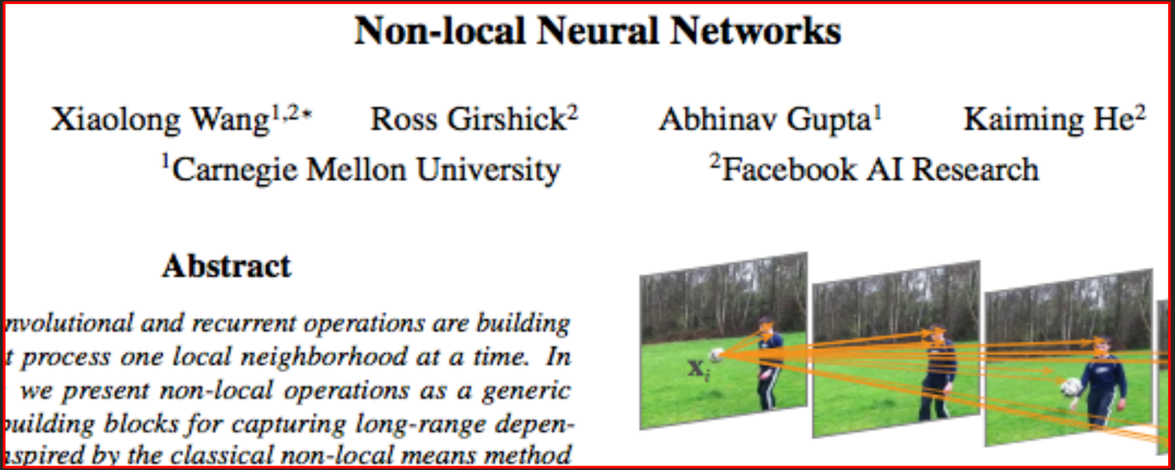
How to better model temporal relation?



How to make model more efficient?

-Remove the dependency on optic flow.

-Sampling of discrete frames



Non-local Neural Networks

Xiaolong Wang[1,2*] Ross Girshick[2] Abhinav Gupta[1] Kaiming He[2]
[1]Carnegie Mellon University [2]Facebook AI Research

Abstract

nvolutional and recurrent operations are building t process one local neighborhood at a time. In we present non-local operations as a generic ouilding blocks for capturing long-range depen- ıspired by the classical non-local means method

https://arxiv.org/pdf/1711.07971.pdf

Future Directions in Activity Recognition

Activity forecasting

What's next?

kiss, hug, highfive

Understanding long videos

Such as movie and TV shows?